DOCUMENT RESUME

ED 169 597 CS 502 490

AUTHOR Steele, Joe M.
TITLE Assessing Speaking and Writing Proficiency Via Samples of Behavior.
PUB DATE Apr 79
NOTE 13p.; Paper presented at the Annual Meeting of the Central States Speech Association (St. Louis, Missouri, April 5-7, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Communication Skills; Daily Living Skills; Educational Research; *Evaluation Methods; Higher Education; *Performance Tests; Speech Skills; *Test Reliability; *Test Validity; *Undergraduate Students; Undergraduate Study; Writing Skills
IDENTIFIERS *College Outcome Measures Project

ABSTRACT

The College Outcome Measures Project (COMP) was established to develop, validate, and implement measures of the knowledge and skills that undergraduate students are expected to acquire and that are important to effective adult functioning. The COMP Measurement Battery yields a "communicating" score based in part on a speaking assessment and a writing assessment, each composed of three role-playing tasks. A high content validity for the scales has been established due to input from communication experts at a diverse group of institutions, the use of standardized role-playing tasks rather than experience-based writing topics, the sensitivity of the assessment to the levels of proficiency normally acquired by undergraduate students, and the establishment of the importance to effective adult functioning of the skills defined by the COMP rating scales. Three studies of interrater reliability for the COMP measures were conducted. In the first study, the correlation between two independent ratings of 74 speaking samples was .83. In the second study, the correlation between two independent ratings of 61 writing samples was .70. In the final study, two raters who had been trained by a refined set of training materials achieved .89 correlation in rating 50 writing samples. The studies show that speaking and writing samples can be measured reliably. (GT)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED169597

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ASSESSING SPEAKING AND WRITING PROFICIENCY VIA SAMPLES OF BEHAVIOR

By

Joe M. Steele
College Outcome Measures Project
The American College Testing Program
P.O. Box 168
Iowa City, Iowa 52243

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Joe M. Steele

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

PRINTED IN U.S.A.

April 1979

Presented at the Central States Speech Association Meeting in St. Louis.

CS502490

ASSESSING SPEAKING AND WRITING PROFICIENCY VIA SAMPLES OF BEHAVIOR

Joe M. Steele
The American College Testing Program

Communicating skills, especially speaking and writing, are held to be extremely important for effective functioning in a variety of adult roles. They are also known to be difficult to develop and troublesome to assess. When I say "known to be," I mean that it is common knowledge--the knowledge of experience. Research studies are rare that establish the relevance of any given definition and its assessment of communication skills to effective functioning in adult roles. This is especially true of objective tests developed as proxy measures of more direct assessment of communication behaviors.

Furthermore, tasks used in the assessment of communication skills often bear little resemblance to real-life tasks faced by adults in various social roles. Hence, for several reasons the validity of existing measures is open to question.

A further difficulty in the assessment of writing and speaking samples is that their evaluation can lead to disagreement among judges as to the levels of proficiency demonstrated. Such disagreement can result in unreliable evaluation of proficiency in communication skills.

This paper seeks to address validity issues as well as reliability issues in the context of a national project to assess various liberal education outcomes of college.

In 1976 the American College Testing Program (ACT) organized the College Outcome Measures Project (COMP) to design, develop, validate, and implement assessment materials to measure the knowledge and skills (1) that undergraduate students are expected to acquire as a result of general education programs, and (2) that are important to effective functioning in adult society. The areas upon which the project focuses are: Communicating, Solving Problems, Clarifying Values, Functioning within Social Institutions, Using Science and Technology, and Using the Arts.

The initial efforts by COMP to develop standardized, content-fair, and cost-effective methods of assessing generic skills in applied contexts have been encouraging. One instrument, the COMP Measurement Battery, uses materials from the adult public domain as stimuli for realistic problems faced in adult roles.

Questions in the Measurement Battery are based on television documentaries, recent magazine articles, ads, short stories, art prints, music, discussions, and newscasts. Participants view these materials and then respond with short written answers, some longer essays in the form of letters, and some oral responses, which are videotaped or audiotaped. The rating procedure used to score these written and oral responses define five levels of proficiency that incorporate both qualitative and quantitative criteria. Guidelines are provided to minimize various sources of bias in rating and faculty evaluators are provided training in applying the rating scales appropriately.

In two test-retest studies of the Measurement Battery, total test score reliabilities of .90 and .89 were achieveed for this six-hour,

open-ended examination. Subscore reliabilities of .66 to .88 were achieved. On the Communicating subscore, the reliability coefficients were .80 and .79. When two of the components of the Communicating subscore, Speaking and Writing, were examined, however, the reliability coefficients were found to range from .37 to .72 (Forrest & Steele 1978). Three studies summarized in this paper were initiated to determine whether improvements in rating procedures plus training of evaluators would raise interrater agreement to acceptable levels.

Assessing Communicating Skills and Knowledge

In the COMP Measurement Battery, communicating is defined broadly as the ability to send and receive information in a variety of modes (written, graphic, oral, numeric, and symbolic), within a variety of settings (one-to-one, in small and large groups), and for a variety of purposes (for example, to inform, to understand, to persuade, and to analyze). Considerable effort has been made to develop an instrument that is content-fair. One must communicate about something, and some people are more knowledgable in some content areas than others. To be fair, an assessment of communicating should cover a broad scope of content. In the Measurement Battery, communicating is defined as the ability to communicate about social institutions, science and technology, and the arts. Two written, two audiotaped and one videotaped stimuli are used in each of the three content areas as the contexts for communicating tasks. Topics for communication about social institutions include governmental agencies, the economy, the family, employment, and volunteer organizations. Technological topics include transportation, housing, energy, food production, and health. Artistic topics include short stories, architecture, music, drama, and painting. The rating procedures require the efforts of college faculty expert in each of the three content areas as well as faculty expert in written and oral communication.

The resulting Communicating score represents a summed judgment of the respondents' ability to read, write, listen, speak, and receive and send information using graphic and numeric representations.

Speaking and Writing Assessment Procedures

The COMP Speaking Assessment is composed of three role-playing tasks for which the respondent has 24 hours to prepare. In Task 1, the respondent is asked four questions about an organization in which he or she is or was a member. A total of three minutes is allowed for response. The student names an organization and, after describing its purpose, elaborates about problems limiting achievement of purposes and changes that could be introduced.

In Task 2, the respondent makes a formal, three minute presentation to a city council committee about use of medical technology at a local hospital, and whether more or less is warranted.

In Task 3, as chairman of an art committee, the respondent makes three one-minute responses to questions from the audience at a public meeting to review a painting selected by the committee.

The evaluation task calls for a holistic judgment of each speaking sample of the degree to which the speaker makes and maintains contact

with the relevant audience, structures a discourse that responds to the questions or points to be addressed, and shows concern for various aspects of delivery. The overriding guideline for the judgment is the degree to which the response represents effective speaking for an adult engaged in these roles.

The COMP Writing Assessment is composed of three role-playing tasks using audio-taped stimuli in the content areas of social sciences, natural sciences and the arts, which require a total of 60 minutes of writing to compose three written communications directed to various audiences. The rating procedure used to score these writing samples requires that a person expert in written communication (e.g., college-level English faculty) apply qualitative rating scales that define five levels of proficiency on three broad criteria to each of the three writing samples for each student. Guidelines are provided to minimize various sources of bias in rating.

For each task, respondents have twenty minutes in which to prepare their written communication to specified audiences. In Task 1, the repondent listens to a taped group discussion of a "Residential Family" and then writes a personal letter to a relative that explains and reflects on an invitation to join this "family." In Task 2, the respondent listens to a taped newscast on the energy shortage and then writes a persuasive letter to the editor of a local newspaper reacting to the newscast and suggesting a local response to the problem. In Task 3, the respondent listens to a tape of two types of music and then writes a memorandum to an employer endorsing one type as background music appropriate for the firm's waiting room.

The evaluation task calls for a holistic judgment of each writing sample of the degree to which the communicator addresses the intended audience, develops and organizes the points called for in the role-playing situation, and makes skillful use of language. The overriding guideline for the judgment is the degree to which the response represents effective writing for an adult engaged in these roles.

In both speaking and writing, equal emphasis is given to rhetorical concerns (organization or discourse) psychological concerns (making contact with and attending to the perspective of the audience), and practical concerns (using vivid language, illustrations, etc. to dramatize and create an effect). Little emphasis is given to formal errors such as misspelled words, punctuation, choppy delivery, etc., unless they noticeably detract from or obscure the message. Three criteria in each rating scale are qualitatively scaled to allow a norm-free judgment of the overall effect.

Faculty evaluators are also provided with the following guidelines to minimize various sources of bias:

1) All students should be rated on one task before the next set of responses is rated.
2) Appearance or penmanship should be disregarded and no credit or penality be given for a position or attitude that you endorse or that you cannot accept.
3) When all students have been rated on a task, the first half dozen should be rerated to assure that interpretation of the scale did not shift after experience was developed in applying the scale.

4) Before beginning to rate the next sample, student response booklets or tapes should be stacked in a different order.
5) Avoid attempting to find out (or be influenced by) the identity, gender, age, grade level, or other characteristics of the students.

Validity Issues

From the beginning, the development of the COMP assessment materials has involved considerable input from faculty experts in communication at a diverse group of institutions participating in the College Outcome Measures Project. This process has assured that the outcomes as defined by the rating scales represent expectations which relevant faculty view as important and which are held in common by a diverse group of postsecondary institutions. Thus a high content validity for the scales has been established.

A second aspect of validity has to do with the nature and design of the tasks themselves. One feature which measures have often lacked in the past has been a standardized task. For example, respondents have sometimes been asked to select a topic, such as "My summer vacation," from a list of topics. The context, other than a broad general topic is entirely defined by the respondent. Such an approach introduces considerable variation into the nature of the response and makes comparisons of student performance difficult. Moreover, the task is not realistic. It does not deal with a situation an adult normally would respond to in writing, is not directed to any audience other than the evaluator, and does not set forth a purpose or outcome the communication should achieve.

COMP has taken a different approach. Rather than selecting a topic that draws on one's private fund of experience, the respondent is asked to play a specific role likely to be encountered by adults. Relevant details are provided as a context in which to respond. The individual is encouraged to use knowledge and skills acquired from any source, but the content needed to respond is information the generally educated college graduate might be expected to know. The COMP assessment materials are designed to measure college exit-level skills in general education. As students may be more knowledgeable in their own areas of specialization, the assessement is made content fair by providing a problem in each of the three broad areas of social sciences, natural sciences, and the arts. Thus, some tasks should provide opportunities to demonstrate proficiency in a content area with which the respondent is familiar.

A third concern related to the validity of the assessment has been to assure that it was sensitive to the levels of proficiency acquired by undergraduate students through normal postsecondary activities. Note that this issue related not only to the sensitivity of the instrument, but to the degree to which the outcomes measured are amenable to instruction.

The combined mean rating of seniors at seven institutions was found to be significantly higher than the mean for a comparable reference group of freshmen, on the Communicating scale and on the subtests for Writing and Speaking, as shown in Table 1. All differences in means between freshmen and seniors are significant beyond the .004 level. Thus, it would appear that the instrument is sensitive to differences in proficiency across groups who would be expected to differ.

See Table 1

A fourth validity concern was to establish that communicating, speaking, and writing, as defined by the COMP rating scales, are important to effective functioning in a variety of relevant adult roles. As part of a study of successful college graduates at the University of Nebraska-Lincoln (UNL), data were obtained on levels of proficiency in in these three areas (Forrest, 1979). Forty-two graduates identified a priori as effective communicators obtained mean scores significantly higher than college seniors on the COMP Communicating and Writing scales. There was no significant difference between senior and adult means on the Speaking scale.

One might assume that a sample of college seniors contains some who are (or will be) effective in communication roles and some who are not. Based on this assumption, one would expect a group of experienced adults, selected because of their effective involvement in communication roles, to score higher than college seniors.

If the graduates had not scored higher than seniors, the data would have raised a question about the validity of these assessments as tapping dimensions important to effective functioning. While the evidence does not prove a necessary relationship, the data in Table 1 suggest that the communication skills other than speaking assessed by COMP are relevant to effective functioning in a variety of adult roles related to communicating. In classifying the adults, no differentiation was made in type of communication role. Further studies will be made of the relevance of Speaking to various adult roles.

Reliability Issues

Three studies of interrater reliability are reported in this paper. In the first two studies two experienced communication faculty were trained and then independently rated 74 sets of speaking samples, and 61 sets of writing samples. In the third study, two experienced communication faculty independently followed a refined set of training materials and then rated 50 sets of writing samples.

The rating task required a person expert in written and oral communication to apply qualitative rating scales that defined five levels of proficiency on three broad criteria to each of the three oral or three written responses. Guidelines for rating were provided to minimize various sources of bias in rating. As a training procedure in the first two studies, the two raters involved in the Speaking and Writing studies were given practice in rating ten student responses to each of the three writing or speaking tasks. The raters were asked to then discuss their ratings with each other and justify any discrepant ratings. They were also given feedback on how others had rated the responses. Differences in level as well as spread of rating were pointed out and strategies to compensate were discussed.

Two approaches to studying interrater agreement were used. First, as the five-point rating was continuous data, a Pearson product-moment correlation was calculated between rater 1 and 2 for ratings of all three student responses combined.

Second, percentages of agreement on a pass/fail basis at each level of proficiency were calculated.

Results

Study 1:

For the Speaking subtest, the correlation between two independent ratings by trained raters was .83 as shown in Table 2. This resulted in extremely accurate ratings of student responses at minimally acceptable and advanced levels of proficiency. On a pass/fail basis, trained raters agreed 99% of the time on performance above or below the lowest standard. At the second level of proficiency, 87% agreement was achieved and at the third level there was 77% agreement. At the highest standard, there was 83% agreement. Thus, it appears that Speaking can be objectively and reliably rated using COMP materials and rating procedures.

See Table 2

Study 2:

For the Writing subtest, the correlation between two independent ratings by trained raters was .70 as shown in Table 3. This was a considerable improvement over the coefficients of agreement of .52 and .43 obtained in the test-retest studies. The accuracy of ratings of student responses at minimally acceptable and advanced levels of proficiency was almost as high as in the speaking study. On a pass/fail basis, trained raters agreed 91% of the time about whether performance fell above or below the lowest standard. At the second level of proficiency, 72% agreement was achieved; at the third level, 71% and at the highest standard 82%. This pattern suggests that the lower correlation obtained in the Writing study may be due to disagreements at intermediate levels of proficiency.

See Table 3

An additional study was initiated to determine whether further refinements in the training materials would produce even higher levels of interrater agreement.

Study 3:

In the third study, two experienced communication faculty independently rated three writing samples per student for 50 freshmen. In addition to the guidelines provided in the first study, raters were given a written discussion of ways to identify various problems of rating and ways to compensate for or correct them. A set of ten sample student responses at many levels of proficiency were selected for each of the three writing tasks. Raters were asked to rate each sample

response and then compare the rating with an agreed-upon standard rating for that sample. The two raters did not meet together but followed these training materials independently and after rating the sample responses each rated the fifty sets of writing samples. Table 4 shows the results for this study.

See Table 4

An interrater reliability coefficient of .89 is quite high, and the percentage of agreement on a pass/fail basis is also quite high. Thus the improvement in training materials appears to greatly enhance rater reliability. Similar refinements in training materials have been developed for the Speaking rating scales. Using these training materials, reliability coefficients for speaking should equal or exceed the levels reported for writing, as a correlation of .83 was obtained in the earlier study.

There remain, however, differences in level of rating as reflected by the mean scores. As long as the interpretation of results relates to the relative ranking of individuals at a single institution or for groups rated by the same evaluator, this difference is of no consequence. Agreement in mean level of performance becomes a concern when groups rated by different judges are compared, or when the same group is rated by different judges over time to examine change scores. For these purposes at present, it would be appropriate to use several judges and assign the mean rating in cases of disagreement. Another series of studies has been initiated to explore ways in which levels of rating might be brought into agreement through application of different rating conventions and ways of compensating for identified patterns of rating level.

Another series of studies (Steele, 1979) has examined the effect on interrater agreement of varying the number of writing samples and the number of raters. Results of these studies indicated that reducing the number of writing samples below three significantly reduced reliability, while increasing the number of samples beyond three or the number of raters beyond two did not seem to greatly enhance reliability. These studies also indicated that the writing rating scales could be applied reliably using analytic rating procedures, as well as the holistic rating judgment discussed in this paper. An analytic rating procedure rates a list of communication characteristics, in this case Audience, Organization, and Language, rather than a single global judgment.

This paper has not addressed the trade-offs of using one method of assesment, such as writing and speaking samples, over another method, such as multiple-choice questions. It would seem that a minimum requirement for objective methods would be to establish their predictive accuracy as proxy measures compared to more direct assessments of the skills. Other papers have addressed some of these issues (Steele, 1979; Huntley, et al, 1979), although they do not adequately treat cost-effectiveness concerns.

Conclusion

These studies demonstrate that speaking and writing samples can be measured reliably. Moreover, measures can be devised that involve realistic tasks assessing ability to apply communication skills. Preliminary evidence is available that some of the skills as defined are relevant to effective functioning in adult roles. Using such measures, normative data can be generated for college freshmen and seniors. These data will be highly relevant in evaluating the effect of the current emphasis on basic skills at the public school level.

A more immediate and practical use for these measures lies in examining and improving college programs which aim to develop speaking and writing skills. Many postsecondary institutions are currently reexamining their general education programs and the meaning of a liberal arts degree, and these data might aid in studying alternative instructional strategies. Using such measures, students could also be screened and more appropriately placed in programs addressing their needs. For institutions wishing to verify the development of communication skills at acceptable levels, measures such as these could be used to assess growth and certify competence. Finally, the assessment and rating procedures offer a model for college faculty to use in developing and reliably rating other open-ended measures of skills and knowledge.

References

Forrest, A. and Steele, J.M. A Test-retest study of the COMP/ACT Measurement Battery, Form II. COMP Annual Report 1978. Iowa City: The American College Testing Program, 1978.

Forrest, A. A study of UNL College of Arts and Sciences graduates. COMP/ACT Technical Report. Iowa City: The American College Testing Program, February 1979.

Huntley, R.M., Schmeiser, C.B., and Stiggins, R.J. The assessment of rhetorical proficiency: The role of objective test and writing samples. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1979.

Steele, J.M. A multiple-choice instrument as a proxy measure for an open-ended assessment of skills. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

TABLE 1
A Cross-sectional Comparison of Performance
in Speaking, Writing and Communication Across
Levels of Educational Development

Level	N	Speaking		Writing		Communicating	
		mean	s.d.	mean	s.d.	mean	s.d.
College Freshmen	131	5.13	2.56	4.56	1.92	26.73	7.55
College Seniors	96	6.39	2.61	5.39	1.73	32.88	8.39
Effectively Functioning Communicators	42	6.55	2.86	7.69	2.26	38.57	10.24
Range		0-12		0-12		0-72	

TABLE 2
Comparisons Between Two Raters who
Rated Three Speaking Samples Per Student for 74 Students

	Rater 1	Rater 2	% of Agreement (Pass/Fail at:
mean	8.16	7.73	Level A = 83%
sd	2.23	2.59	Level B = 77%
r	.83		Level C = 87%
			Level D = 99%

TABLE 3
Comparisons Between Raters who Rated
Three Writing Samples Per Student for 61 Students

	Rater 1		Rater 2	% of Agreement (Pass/Fail) at:	
Mean	7.23		5.52	Level A	82%
sd	2.67		2.24	Level B	71%
r		.70		Level C	72%
				Level D	91%

TABLE 4
Comparisons Between Two Trained Raters who
Rated Three Writing Samples Per Student for 50 Students

	Rater 1		Rater 2	% of Agreement (Pass/Fail) at:	
Mean	5.84		4.80	Level A	94%
sd	2.71		2.42	Level B	85%
r		.89		Level C	79%
				Level D	84%